A Study of the Tovota Production System
From an Industrial Engineering Viewpoint

by Shigeo Shingo

1989

Dr. Shigeo Shingo and Mr. Taiichi Ohno created the backbone of the industrial revolution known as the Just-In-Time system, the powerful effects of which have altered the international economic order. Dr. Shingo is the teacher and Mr. Ohno is the manager.

Dr. Shingo's study explains the philosophy behind the Toyota Production System, provides additional information where required, criticizes weaknesses, gives credit where it is due, and highlight the system's important aspects.

Shingo has developed the following conclusions from his study:

- Elimination of the waste of over-production cannot be achieved without SMED (single minute exchange of die)
- Shortened cycle times demand small lot production (SMED is crucial here as well)
- SMED must be achieved if we want to be able to respond to changes in consumer demand

While Dr. Shingo was developing SMED, Mr. Ohno, as the executive managing director of Toyota, was realizing that there was a close connection between the principle aspects of the Toyota Production System and the SMED system. In his pursuit of the ideal production system, he became convinced that drastic reductions in set-up times were essential. This conviction prompted him to demand that his people achieve three minute die changeovers. Fortunately, his demand and Shingo's ideas about the SMED system that were in the back of his mind, coupled with Shingo's keen interest in set-up reductions all came together in a timely manner.

Many people, when they begin to implement the SMED system, believe that know-how is important. Shingo believes that to succeed you must know-why as well. With know-why, you can understand why you have to do what you are doing, and hence will be able to cope with changing situations. If you only acquire know-how, you may not be able to implement SMED effectively in your own operations where production characteristics most likely differ from those at Toyota.

Shingo believes that the significant principle and unique feature of the Toyota Production System lies in the following:

- In order to eliminate inventory (previously thought to be a necessary evil) various basic factors must be thoroughly explored and improved.
- Relentless efforts must be made to cut man-power costs.
- Total elimination of waste is the basic principle of the Toyota system.

In the Toyota system, there are two words that have very specific meanings: process and operation. Process is the course by which material is transformed into product. This consists of processing, inspection, transport and storage. Operations are the actions performed on the material by machines and workers.

All production, whether carried out in a factory or the office, must be understood as a functional network of process and operation. A process transforms material into product. It is accomplished through a series of operations. A process is a flow of material in time and space; a transformation from raw material to semi-processed components to finished product. Operations are the work or actions performed to accomplish this transformation; the interaction and flow of equipment and operators in time and space.

Process analysis examines the flow of material or product. Operation analysis examines the work performed on products by worker and machine. To make fundamental changes in process, we must distinguish product flow (process) from work flow (operation) and analyze them separately.

Improving Process

In order to maximize production efficiency, thoroughly analyze and improve process before attempting to improve operations. There are four distinct process elements in the flow of raw materials into products.

- Processing - a physical change to the material or its quality
- Inspection - comparison with an established standard
- Transportation - the movement of material or products; a change in its position
- Delay - a period of time during which no processing, inspection or transport occurs

There are two types of delay:

- Process delay - an entire lot waits while the previous lot is processed, inspected or moved.
- Lot delay - while one piece is processed, the others wait. They wait either to be processed or for the rest of the lot to be done.

There are two types of process improvements. Value engineering asks "How can this product be redesigned to maintain quality while reducing

manufacturing cost?" The second stage of process improvement ask the question "How can the manufacturing of this product be improved?"

Improving Inspection

Traditional processes use judgment inspection. This simply means distinguishing defective from non-defective products. Improving judgment inspection merely increases the reliability of the inspection process, but has no effect on the defect rate. improving the process will reduce inspection errors.

To actually reduce the defect rate, informative inspection must be done. Processing must be informed whenever a defect is discovered, so that steps can be taken to correct the processing method or condition and prevent an occurrence. Rather than doing a "post-mortem", a symptom or defect is identified and "treated." Judgment inspections discover defects, while informative inspections reduce them. The purpose of inspection must be prevention. Quality can be assured reasonably only when it is built in at the process, and when inspection provides immediate, accurate feedback at the source of defects, rather than at the end of the process.

Types of informative inspection

Self inspection and successive inspection

In a self inspection, the worker inspects the product he/she processes. Two drawbacks are that the worker may compromise judgments and accept items that ought to be rejected or make inspection errors unintentionally.

Successive inspection provides both objectivity and immediate feedback, on the other hand. In successive inspection, workers inspect products passed along from the previous operation before processing them themselves. On average, and 80-90% reduction in the number of defects can be achieved within one month of adopting the successive inspection system.

Enhanced self inspection

This provides the fastest feedback. Self inspection can be enhanced with the use of devices that automatically detect defects or inadvertent mistakes. Such systems give the individual worker immediate feedback, achieve 100% inspection, and prevent defects. Physical detection devices are called poka-yoke or "mistake-proofing" devices.

Source inspection

Source inspection prevents defects by controlling the conditions which influence quality at their source. Vertical source inspection traces

problems back through the process flow to identify and control external conditions that affect quality. Horizontal source inspections identify and control conditions within an operation that affect quality.

Poka-yoke Inspection Methods

Poka-yoke achieves 100% inspection through mechanical or physical control. Poka-yoke can either be used as a control or a warning. As a control it stops the process so the problem can be corrected. As a warning, a buzer or flashing lamp alerts the worker to a problem that is occurring.

The three types of control poka-yoke include:

- Contact method - identify defects by whether or not contact is established between the device and some feature of the product's shape or dimension

- Fixed value method - determines whether a given number of movements have been made

- Motion step method - determines whether the established steps or motions of a procedure are followed

The first step in choosing and adopting effective quality control methods is to identify the inspection system that best satisfies the requirements of a particular process. Only after deciding on whether control or warning poka-yoke is necessary should the actual type of device be considered.

Transport Improvement

Transport is a cost that does not add value to a product. Most people try to improve transport by using forklifts, conveyors, etc. which actually only improves the work of the transport. Real improvement eliminates the transport function as much as possible. This involves improving the layout of process.

Transport improvement and transport operation improvement are two distinctly different problems. Only after opportunities for layout improvement have been exhausted should the unavoidable transport work that remains be improved through mechanization.

Eliminating Storage

There are two type of delays relating to storage:

- Storage between processes (process delays)
- Storage for lot size (lot delays)

Eliminating process delays

Process delay refers to both lots of unprocessed items waiting to be processed and accumulated excess inventory that sits waiting to be processed or delivered. There are three types of accumulations between processes:

- E storage - resulting from unbalanced flow between processes (engineering)
- C storage - buffer or cushion stock to avoid delay in subsequent processes due to machine breakdowns or rejects (control)
- S storage - safety stock; overproduction beyond what is required for current control purposes

Eliminating E storage

This can be eliminated through leveling quantities, which refers to balancing flow between high and low capacity processes. This may require not running high-capacity machines at 100% capacity, in order to match flow with lower capacity machines that are already running at 100%. It may also mean replacing high capacity machines with lower capacity machines, in order to better match the flow of the whole process.

E storage can also be eliminated through synchronization. This simply refers to efficient production scheduling so that once quantities are leveled (output is matched), processes are scheduled with those ahead and behind in the overall process, so that inventories need not pile up do to scheduling conflicts.

Eliminating C storage

Cushion stocks compensate for machine breakdowns, defective products, downtime for tool and die change, sudden changes in production scheduling, etc. When these issues are addressed through new processes, cushion stocks can be avoided. Steps should include such things as:

- determining the cause of machine failure at the time it occurs, even if it means shutting down the line temporarily

- better inspection processes to avoid defective parts

- implementing SMED to eliminate long set-up times

- running smaller batch sizes to allow for quick changes in production plans

Eliminating S storage

"Safety stock" is in place to guard against delivery delays, scheduling errors, indefinite production schedules, etc. Shingo recommends keeping a small controlled stock that is only used when the daily or hourly scheduled delivery fails or falls behind. The safety stock can then be replenished when the scheduled materials arrive, but the supply of materials due for the process go directly to the line, rather than normally going into storage first. This is the essence of the just-in-time method.

Eliminating lot delays

While lots are processed, the entire lot, except for the one piece being processed, is in storage. The greatest reduction in production time can be achieved when transport lot sizes are reduced to just one; the piece that was just worked on. These transportation changes can be accomplished through changes in layout of the entire operation flow using conveyors which result in shorter production cycles and decreases transport man-hours. Using SMED (single-minute exchange of dies), set up time is decreased so large lot sizes are no longer necessary to achieve machine operating efficiencies.

Improving Operations All operations can generally be classified as:

- Set up operations - preparation
- Principal operations - performance
- Margin allowances - machine breaks
- Personal allowances - worker breaks

Improving set up

The most effective way to improve set up operations is using SMED. There are two types of set ups; internal, which can only be performed when the machine is stopped, and external, which can be performed while the machine is running. SMED focuses on changing internal to external wherever possible.

SMED techniques as developed by Dr. Shingo are: Separate internal from external operations

- Convert as many internal operations as possible to external ones
- Standardize functions, not shape - i.e. use same clamps for all set ups
- Use functional clamps or eliminate fasteners altogether - i.e. use wedges, cams, and other one-touch methods to fasten dies

- Use mediated jigs - centering can be done as an external set up if all jigs are standardized Adopt parallel operations - if a set up requires work on 2 sides of the machine, use two people for the set up, thereby cutting set up time by more than half

- Eliminate adjustments - use finite number of limit switches to determine settings instead of turning screws, which have unlimited "settings"; the
- Least Common Multiple theory makes the number of settings finite and unvarying

- Mechanization - only use after the previous 7 steps have been exhausted

Improving principal operations

The easiest way to improve principal operations is to separate the worker from the machine. This involves the "one worker, many process" theory. Whereas workers must be paid indefinitely, machines depreciate and eventually can be operated for "free." This concept makes it much more productive to have one worker attend 5-6 machines, even if some of the machines are temporarily idle. Cost reduction is more important than high machine operating rates.

Improving margin allowances

Processes should be examined for automation improvement such as automatic lubrication or using oil-impregnated metals, thereby eliminating the need for time consuming "machine maintenance." Workshop allowance should also be examined for streamlining, to include such processes as automatic product storage (thereby not needing a worker to manually move products to storage).

Developing Non-stock Production

A principal feature of the Toyota Production System is an emphasis on non-stock production. There are two types of stock:

- Naturally occurring stock which accumulates from incorrect forecasts and overproduction

- "Necessary" stock which is planned in advance to account for fluctuations in the order cycle, delays in production and machine breakdowns

It should be understood that all stock is wasteful and unprofitable. Conditions that create stock should be corrected so stock naturally reduces. Examples are reducing production cycle time, eliminating breakdowns and delays by determining their causes when they occur and reducing set up times.

Principles of TPS

Most people think that TPS is the kanban system. TPS is actually a system for the elimination of waste. Kanban is just a means of achieving just-in-time.

Defining basic principles

Waste of overproduction can be categorized as either quantitative or early. Quantitative overproduction results from making more than is needed. Early overproduction results from making product before it is needed. Neither is desired in TPS.

In Japanese, the word for just-in-time actually means "timely", "well-timed" or "just-on-time." Just-in-time is a system where each process is supplied with the required items, in the required quantity, at the required time. It does not allow for early or late delivery or accumulation of product.

Separation of worker from machine refers to increasing production efficiency through lower machine operating rations, and promoting more effective use of human resources.

Perform an appendectomy refers to shutting down machines if necessary in order to determine exactly what defect or problem has been detected. It comes from the idea that using an ice pack when a person has appendicitis eases the pain, but doesn't cure the problem. TPS says that the only reason to stop the line is to ensure it won't have to happen again. The commitment to problem solving is 100%.

Fundamentals of Production Control

The first fundamental is using a non-cost principle. Instead of the manufacturer trying to determining selling price by using a formula of cost plus profit, TPS uses a principle of selling price minus cost equals profit. It understands that the market determines price, not the seller.

The second fundamental is the difference between value adding operations and non-value adding operations. Only processing adds value. All other operations, such as inspection and transport, are non-value adding. TPS works to eliminate as many non-value adding operations as possible. There is no such thing as an "improvement" on a waste.

The third fundamental is the idea of problem solving using the "5 whys." All problems are solved, all defects are addressed, by asking why five times, until the root cause of the problem is determined. The problem is not usually that the machine shut down, but rather that some other part of the process is not the best it could be and caused a defect to occur.

Mass Production and Large Lot Production

While manufacturers may not always have the option of choosing between small, medium or mass production, they do always have the option to choose between small or large lots. Small lots are always preferred, even in mass production, because they reduce excess inventory.

Large lot production is speculative. TPS is based on confirmed orders and geared toward fast delivery of a wide variety of models. TPS sets capacity at minimum demand level and levels fluctuations in demand accordingly.

Characteristics of order-based production

- Overtime work
- Excess machine capacity run with temporary workers
- Order based production for season demand - prepare for 70% of expected demand and use direct orders to drive excess capacity production
- Order delivery period must be greater than production cycle
- Speedy delivery
- Strong market research
- Production planning driven by order-based demand

Mechanics of TPS

Schedule control and load control are two important concepts. Schedule control ensures that the product is made on time. Load control ensures that the product can in fact actually be made.

TPS uses three schedules. The master schedule is long term (one year), the intermediate schedule is fine tuned one to two months in advance, and the detailed schedule is the practical production schedule for a day or a week.

TPS runs on seven principles for shortening the production cycle. Shortening the production cycle relies on eliminating wastes, which we have stated is the key principle of TPS.

7 principles to shorten the production cycle

Reduce process delays - level quantities and synchronize the flow of product through the plant

Reduce lot delays - keep work-in-process moving using a one-piece flow method; lots are single products, not batches

Reduce production time

Employ layout improvements and the full work control system - alter layout so little or no transport has to be done between processes; the plant is laid out in a single process line Synchronize operations and absorb deviations - workers assist each other when one slows down, so that inventories don't back up between operations; processes are synchronized to flow directly from one to the next without delay

Establish "tact" time - the exact time required to process one piece of product Ensure product flow between process

100% inspection

Instituting poka-yoke inspection ensures that no defective products are passed down the line to the next process. Poka-yoke refers to use of a jig or fixture that helps determine 100% acceptable product. For example, if a piece doesn't fit properly into the jig, it is not made to specification and is therefore defective. Poka-yoke simply refers to "fool-proofing" a process. Either the piece will not fit, the machine will not start, or if the correct sequence is not followed the machine will notify the operator.

Leveling and Load Averaging

Leveling refers to matching the quantity produced to the quantity taken from the process that precedes it. Timing and volume are critical here.

Load averaging refers to determining production quantities for the month, and then breaking them down into smaller production batches. For example, if 3,000 units of car A are needed this month, 1,000 will made in the first 10 days of the month, 1,000 made the second 10 days, and 1,000 made the third 10 days. Mixed production refers to assembling vehicles along the line in a variety of combinations, which are determined by how many are to be made during the next 10 days. Ideally, load averaging can be broken down to how many of each type of car are to be created each day of the month, or hour of the day, and then they are "mixed" with the other models being made along their individual schedules.

Mixed production can only be accomplished, though, through small set ups, which can be done in a few minutes, instead of a few hours.

The Nagara System

The fundamental concept behind the Nagara system is the breakdown of shop divisions. It denies the idea that forging must be done in a forge, painting in the painting shop, etc. In the Nagara system, on worker performs more than one process at a time, by automating as many processes as possible. A worker can then start an automated process, go start another one, and come back to unload and reload the first one. An automated press can then work next to a person who is spot welding the pieces that come directly out of the press. The process further
streamlines the idea of continuous process, regardless of function. (i.e. the lathes are no longer all together, but placed amongst the process wherever they fall in the stream of operations)

Standard operations

In TPS, all operations are standardized. A standard operation is broken down into cycle time, work sequence and standard inventory. Cycle time is how long it takes to make one piece. Work sequence is the order in which a worker must process an item. Standard inventory is the minimum number of intra-process pieces needed for operations to proceed, including pieces in machines.

Workers write out their standard operations on a "standard work sheet." Describing operations on paper allows for objective observations and facilitates operation improvements. Standard work sheets should continuously be revised as improvements are made.

From standard work sheets, standard operating charts are developed. These are used to train new workers in processing. The new workers can continue to refer to the chart until the new process is learned.

There are different types of standard operating charts.

- Capacity charts by part - record order of processes, machine numbers, basic times, processing capacity, etc.

- Standard task combination sheets - order in which individual worker's operations take place

- Task manuals - procedures for elements of operations requiring special attention, i.e. tool changing

- Task instruction manuals - for training workers; guidelines for correctly teaching. standard operations

- Standard operating sheets - equipment layout diagrams from Task Instruction Manual; displayed on the shop floor; includes cycle times, order of operations, safety and quality checks

Manpower cost reductions

Since idle workers cost more than idle machines, TPS always looks to improve how it uses manpower. This included improving human motions, improving machine motions (in order to use fewer human motions), and mechanizing human motions, such as having a machine attach and remove work pieces from another machine.

TPS is careful not to mistake qualitative reductions in manpower with quantitative reductions. For example, using mechanization to make work easier may save a worker time. However, if the worker is now idle while the machine does the work, there is no actual savings. The process is still taking the same amount of worker time. Labor saving improvements are not confused with worker saving Improvements.

Wherever possible, machine layout and worker layout are determined independently. Machines are laid out outside of a chosen pattern, such as a U-shaped layout, and the workers are then stationed inside. This reduces isolation of workers and facilitates mutual assistance.

When a process line is laid out continuously, instead of breaking machines apart by function (all the lathes in one place, all the presses in one place), the line can increase the number of machines that one worker operates. This is referred to as multi-process handling. One worker follows one piece through a number of processes. Machine "wait" time may increase, but "human time" decreases, thereby costing less.

Structure of TPS

Basic features of TPS

- Targets cost reduction via the thorough elimination of waste

- Eliminates overproduction through the notion of non-stock and achieves labor cost reduction via minimal manpower - the two aspects of production in which the most waste occurs

- Reduces production cycles drastically through the use of the SMED system to achieve non-stock by carrying out small lot production, equalization, synchronization, and one piece flows

- Thinks of demand in terms of order-based production. To attain this under non-stock conditions, looks at all problems from a fundamentals-oriented perspective

- Adheres consistently to the idea that the quantity produced should be the quantity ordered

Kanban

The kanban system is a means to control just-in-time supply and "autonomation" (automation with a human touch). The kanban system works hand-in-hand with the "order-point method."

The order-point method is a control technique used to carry out optimum ordering in repetitive production processes. It is a technique/formula used to lower inventories using smaller and smaller lots, thereby increasing the frequency of delivery of materials. Because materials are delivered more often, new strategies are needed to deal with the amount of increased transport, in order to ensure that excessive transport waste is not created.

Kanban is a means of visual control, used to keep the supply system going. The theory behind kanban is that only what is used is replenished. By only creating what was taken, it creates a "pull" of inventory through the system, rather than "pushing" material through that was created without need.

In the traditional kanban system, a kanban fulfills three main functions.

- Identification tag - indicates what the product is

- Job instruction tag - indicates what should be made, for how long and in what quantities

- Transfer tag - indicates from/to where the item should be transported.

In TPS, there are only two tags which fulfill the three above functions.

- Work-in-process tag - serves as identification and job instruction tags

- Withdrawal tag - serves at identification and transfer tags

The two main features of a kanban system. First, kanban are used repeatedly. Secondly, the number of kanban are restricted, which limits product flow, which in turn eliminates waste and holds stock to a minimum. The goal of a kanban system (and TPS) is to minimize the number of kanban in the system.

Peripheral Issues

There are seven types of waste in a production system.

- Overproduction
- Delay
- Transport
- Processing
- Inventory
- Wasted motion

Making defective parts The goal of TPS is to eliminate as many of these as possible.

Just-In-Time production

Just-in-time production is controlled largely by the relationship between the order-to-delivery cycle (D) - how long it takes to receive payment from the time the order is taken - and the production cycle (P) - how long it takes to make the product. As long as D is longer than P, production beginning after a firm order is received will still be on time without having to generate inventory. If D is shorter than P, inventory must be generated and kept on hand in order to shorten the production schedule. Just-in-time only works as long as D is greater than P.

The Future Course of TPS

In order to continue to reduce set up times, TPS must move from SMED (single-minute exchange of dies) to OTED (one-touch exchange of dies). This will be accomplished using the least common multiple method, automating changeovers, using finite limit switches for adjustments, and determining identical and different parts. By determining which parts of dies are identical and which parts are different, simple methods can be devised to switch only the different parts.

The ultimate goal, of course, is no-touch set ups. The fastest way to change something is to change nothing, so making parts in sets of two will eliminate a set up between every other part.

Further development of the continuous flow system will also eliminate wastes. Currently, assembly and processing are done in a continuous flow. TPS will continue to try to add processes further upstream to the flow, such as welding,

forging and casting processes. Nagara combats this with the idea that like functions need not be done in the same place and tries to move functions to the place in the process where they are needed.

Cutting labor costs is a continuous process in TPS. The process of cutting labor costs includes (in order of least expensive improvements first):

- improve human work motions
- integrate margin allowances
- shift human work to machines make machines that can detect abnormal situations make machines that can detect and RESPOND to abnormalities (can fix themselves)

Dealing with Decreases in Demand

TPS encounters decreases in demand, just like mass production does. Where mass production tends to lay off workers in slow periods and hire workers in growth periods, TPS includes the concept of hiring a worker for life. Therefore, in slow growth periods, TPS decreases the number of people on the line by increasing the number of machines each worker oversees. They then take the extra workers from the line and use them for unscheduled repairs or maintenance of machines, has them practice set ups so they will go faster when production speeds up again, or uses them for general cleaning and maintenance. Laying off workers is not an option, though. The central idea in TPS is not hiring too many workers in the first place.

Implementing TPS

It is a mistake to try to merely imitate TPS. External features of the system can only be successfully applied based on a thorough understanding of the principles involved. The principle of increasing profit through eliminating waste must be embraced by management first, so that when the line needs to shut down to determine the root cause of problem, thereby eliminating that waste a second time, management will support the decision.

Shingo recommends using "cushion" stock when first changing over to TPS. Current stock can be used as the cushion, but the amount should be sealed (set at current levels). As delays occur during the learning and trial periods, stock can be pulled, but should be replaced the next day. The amount of necessary stock will lessen as waste is eliminated and lots get smaller (set up times are getting shorter). He cautions against eliminating stock too quickly, though. This usually ends up stopping the line due to shortage as well as due to problem-solving.